Fronthaul-Aware Scheduling Strategies for Next Generation RANs

Sandra Lagén*, Xavier Gelabert[†], Lorenza Giupponi*, Andreas Hansson[†]

*Centre Tecnològic de Telecomunicacions de Catalunya (CTTC/CERCA), Castelldefels, Barcelona, Spain

[†]Huawei Technologies Sweden AB, Kista, Sweden

{sandra.lagen, lorenza.giupponi}@cttc.es, {xavier.gelabert, andreas.hansson}@huawei.com

Abstract—Next generation Radio Access Networks (RANs) consider virtualized architectures in which base station functions are distributed in different logical nodes, connected through fronthaul (FH) links. To reduce the FH deployment costs and the required FH capacity, operators may install a single FH link shared among multiple cells and exploit key enabling techniques, such as modulation compression, to decrease the data rates over the FH. In shared FH capacity scenarios, it is essential to provide efficient methods to control and optimize the FH resources' utilization with limited impact on the air interface performance. In this paper, we propose and analyze different fronthaul-aware scheduling strategies, leveraging on modulation compression, for multi-cell multi-user scenarios with a shared FH link across multiple cells. We consider holistic approaches based on packet dropping at the PHY layer and postponing of MAC scheduling decisions, combined with the reduction of the modulation order per cell. Additionally, we propose optimization methods for resource allocation and dynamic modulation compression, in which the modulation order and resource block assignment is dynamically optimized per user and slot. We finally evaluate the proposed FH-aware scheduling methods over an end-to-end dynamic 5G NR system-level simulator based on ns-3.

Index Terms—next generation RAN architectures, C-RAN, modulation compression, fronthaul-aware MAC scheduling, optimization.

I. INTRODUCTION

Next-generation RANs need to deal with several challenges to provide high spectral efficiency, low power consumption, resource pooling, scalability, and cross-layer interworking. For that, the 3GPP and Open-RAN (O-RAN) initiatives envision hybrid Centralized-RAN (C-RAN) architectures [1], in which the baseband processing of a single cell is disaggregated between different physical baseband entities, while at the same time parts of the baseband processing of multiple cells are collocated in centralized locations. In particular, a 3GPP and O-RAN 5G New Radio (NR) base station (i.e., gNB) consists of a Centralized Unit (CU), one or more Distributed Units (DUs) connected to the CU via a midhaul interface using the so-called Option 2 functional split (PDCP-RLC split) [2], and multiple Radio Units (RUs) that are connected to one DU each through a fronthaul (FH) interface using Option 7.2x functional split (intra-PHY split) [3]. When deploying C-RAN architectures, the main obstacles are tight FH capacity and latency requirements [4]. Such FH requirements are further

accentuated in 5G NR by using wider channel bandwidths, massive antennas, higher modulation orders, and a larger number of aggregated carriers [5], which contribute to increasing the required FH capacity [6].

On the one hand, to reduce the required FH capacity, FH compression schemes can be used [1]. Among the different techniques envisioned by O-RAN [3], modulation compression is considered a promising option because it allows a dramatic reduction of the required FH capacity without degradation of the signal quality sent over the FH interface and without the need for complex algorithms/schemes [6]. Basically, modulation compression achieves FH capacity reduction by means of decreasing the IQ bitwidth¹ sent over the FH interface based on the modulation order used over the air interface. On the other hand, to reduce the deployment and operational costs, shared FH links across multiple cells are of high interest for mobile network operators [7]. Such scenarios are particularly challenging from the technical point of view, as they result in a shared FH capacity utilization across multiple cells. In essence, given a particular C-RAN architecture, with specific functional splits and a fixed FH topology, it is of utmost importance to provide efficient methods for 5G NR to control and optimize the utilization of FH resources with limited (or no) impact on the air interface performance. This calls for system wide designs and optimization solutions, whose evaluations need to be based on dynamic, end-to-end, multi-cell simulations. In [8], semi-static modulation compression methods are proposed to optimize the maximum modulation order that is permitted per cell at the long-term, based on cells' statistics, considering a shared FH interface of limited capacity. Differently, in this paper, we focus on dynamic modulation compression methods, to adjust the scheduling and modulation compression decisions to the available FH capacity on a per slot and per user basis. We present, analyze, and optimize different dynamic FH-aware scheduling methods, including holistic methods and optimization methods for resource allocation and modulation compression, in multi-cell scenarios with shared FH interface. Regarding the numerical evaluation, we assess the proposed strategies using a dynamic multi-cell 5G system-level simulator based on ns-3 [9].

The paper is organized as follows. Sec. II describes the system model assumptions, including the analytical charac-

This work was partially funded by Spanish MINECO grant TEC2017-88373-R (5G-REFINE), Generalitat de Catalunya grant 2017 SGR 1195, and Huawei Technologies Sweden AB.

¹The number of bits used to represent an IQ sample.

terization the FH and air interfaces. Then, Sec. III presents baseline holistic FH-aware scheduling options and dynamic optimization solutions, considering a shared FH capacity constraint. Finally, we describe the simulation scenario and assess the dynamic modulation compression performance in Sec. IV. Sec. V provides some concluding remarks.

II. SYSTEM MODEL

The system model is composed of N cells, whereby each cell serves K_n users (UEs) ($n = 1, \dots, N$), hence totalling $K = \sum_{n=1}^N K_n$ UEs in the system. We assume a hybrid C-RAN architecture with centralized, distributed, and radio units, implementing functional split Option 2 for the CU-DU split and intra-PHY functional split 7.2x for the DU-RU split, as specified by 3GPP [2] and O-RAN [3]. For the CU/DU/RU deployment, we consider the so-called Scenario B, prioritized by O-RAN [10], and in which CUs and DUs of all cells are collocated in a centralized entity (in an edge or regional cloud), while the RUs are located in proprietary cell sites. This way, the high-PHY, MAC and above processing of all the cells are implemented together in the centralized entity, while low-PHY and RF processing of each cell are placed in the RU of each site [11]. A low layer FH interface is used to interconnect RUs with DUs [3]. We assume a star FH topology so that the N RUs share the same FH link, being characterized by a limited capacity of C bits/s. Fig. 1 shows the considered deployment.

We focus on the downlink and adopt the so-called modulation compression technique to reduce the required FH capacity in multi-cell scenarios with shared FH interface. Modulation compression allows lossless IQ data compression (see e.g. [3, §A.5]) whereby modulated data symbols can be represented by a limited number of IQ bits, i.e., the number of bits needed to represent the employed constellation in a data block [6]. In particular, we focus on dynamic FH-aware scheduling methods in which the modulation compression is dynamically optimized per cell and per UE, on a slot basis, considering the actual performances of the FH and air interfaces.

In the shared FH interface, the *fronthaul throughput* that is needed through all the cells ($n = 1, \dots, N$) using functional split 7.2x in a specific slot t (S_t^{FH}) is:

$$S_t^{\text{FH}} = \sum_{n=1}^{N_t^a} \sum_{k=1}^{K_{n,t}^a} \frac{1}{T} (N^{\text{sc}} N^{\text{os}} N_{n,k,t}^{\text{rb}} M_{n,k,t} + \mathcal{O}_{\text{mac}} + \mathcal{O}_{\text{mc}}), \quad (1)$$

where $N_t^a \leq N$ is the total number of cells that are active and share the FH interface in slot t , $K_{n,t}^a \leq K_n$ is the number of UEs being actually served by the n -th cell in slot t , $M_{n,k,t}$ is the modulation order assigned to the k -th UE in the n -th cell at the t -th slot, $N_{n,k,t}^{\text{rb}}$ is the number of Resource Blocks (RBs) allocated for data transmission towards the k -th UE in n -th cell at the t -th slot, N^{os} is the number of OFDM symbols within a slot available for downlink data transmission (which is lower than 14 in 5G NR), N^{sc} is the number of subcarriers within a RB (which is equal to 12 in 5G NR), T is the slot length (in seconds), $\mathcal{O}_{\text{mac}}$ is the MAC information overhead per UE (in bits) needed to implement intra-PHY splits (including RB

Fig. 1. Deployment scenario composed of N cells, each serving K_n UEs, with shared FH interfacing between distributed and radio units (DUs and RUs) with a star FH topology. CUs and DUs are placed together in a centralized entity, and each RU is located in a cell site.

assignment information, antenna configuration, beamforming vectors, etc.), and $\mathcal{O}_{\text{mc}}$ is the signaling overhead (in bits) that accounts for the notification through the FH towards each RU of the modulation order in use for each UE. In (1), single MIMO layer per UE is assumed.

For the air interface, we consider the *per-user throughput* as a metric of interest. The throughput for the k -th UE in the n -th cell at slot t ($S_{n,k,t}^{\text{AI}}$) is given by:

$$S_{n,k,t}^{\text{AI}} = \frac{1}{T} (N^{\text{sc}} N^{\text{os}} N_{n,k,t}^{\text{rb}} M_{n,k,t}). \quad (2)$$

From previous expressions in Eq. (1) and Eq. (2), it is evident that the scheduling decision (i.e., scheduling of UEs, resource allocation, and modulation order assignment) creates an air-FH trade-off that needs to be optimized appropriately. That is, e.g., increasing the modulation order of a specific UE in a slot, increases the required FH throughput for that slot but also leads to a higher per-user throughput on the air interface. The trade-off is further accentuated in multi-cell scenarios, when multiple cells share the FH capacity (as reflected in Fig. 1) and potentially the air interface channel. In those multi-cell settings, suboptimal scheduling decisions for each of the cells, from a system-wide perspective, may lead to situations in which certain cells experience congestion (with low per-user throughputs) while other cells have high spectral efficiency and unused spectrum resources (with high per-user throughputs). Cells in the latter condition could thus reduce the assigned modulation orders, which would spare some FH capacity in advantage to cells with more challenging conditions, thus benefiting from a fair quality-of-service across all the UEs in the network. Having this in mind, efficient centralized and dynamic modulation compression scheduling methods for FH compression control should exploit the shared FH capacity while providing quality-of-service guarantees across all UEs.

III. FH-AWARE SCHEDULING STRATEGIES

Typically, $M_{n,k,t}$ (and so, also the Modulation Coding Scheme (MCS)) for the downlink is set based on the Channel Quality Indicator (CQI) reported by the UE. Based on $M_{n,k,t}$ and the RLC buffer level, the MAC scheduler determines the

number of RBs that will be assigned to each UE ($N_{n,k,t}^{\text{rb}}$). For that, the MAC scheduler follows certain pre-established scheduling rules (e.g., proportional fair or round robin, among others). Herein, to meet the shared FH capacity constraint, we start by introducing two baseline FH-aware scheduling methods, namely: “dropping” and “postponing”, which are easily implementable, as they do not need modifications in the traditional MAC scheduler operation. Subsequently, enhancing the aforementioned schemes, we propose centralized FH-aware scheduling methods that optimize the resource allocation (number of RBs, $N_{n,k,t}^{\text{rb}}$) and modulation compression (modulation order, $M_{n,k,t}$) for all UEs and cells, assuming that all UEs with data on a particular slot have to be scheduled.

A. Dropping packets at PHY

The first baseline FH-aware scheduling option is to implement packet “dropping” at the PHY layer. That is, assume that the modulation order ($M_{n,k,t}$) and the RB assignment ($N_{n,k,t}^{\text{rb}}$) are determined per UE as usual, i.e., disregarding the FH capacity constraint, and the MAC Packet Data Units (PDUs) are prepared and are sent to the PHY layer, as usual. Then, the high-PHY layer of each cell drops the MAC PDUs (including the related control and data) that cannot fit in the available FH capacity (C) by considering the MAC PDUs across all the cells. This implies dropping the full MAC PDU packet(s). Noteworthy, as the UE will not receive any control nor data, no Hybrid Automatic Repeat Request-Acknowledgment (HARQ-ACK) feedback will be generated and, therefore, after a certain timing for HARQ-ACK feedback reception at the gNB has passed, the full HARQ process related to such a drop will be erased. This baseline option acts as a centralized logic that considers the usual RBs/MCS assignment at the MAC scheduler and determines which data (new data and/or HARQ retransmissions) is dropped. The dropping decision is implemented at the high-PHY layer in the DU.

B. Postponing MAC scheduling decisions

The second baseline FH-aware scheduling option is to implement dropping only of the MAC scheduling decision, i.e., not dropping data but “postponing” its transmission/retransmission. In this way, the MAC layer will drop the scheduling decision generated by the MAC scheduler, but data will remain in the RLC queues (new data) or the HARQ buffers (retransmission data) without data dropping. In this case, we assume that the modulation order ($M_{n,k,t}$) and the RB assignment ($N_{n,k,t}^{\text{rb}}$) are determined per UE as usual, disregarding the FH capacity constraint, and then the MAC scheduling decisions are discarded if the assignment does not fit in the available FH capacity limit (C), by considering the already scheduled allocations across all the cells. So, this baseline option acts as a centralized scheduling logic that considers the usual RBs/MCS assignment at the MAC scheduler and determines which scheduling decisions (of new data and/or HARQ retransmissions) are discarded, for which its actual transmission is postponed. The postponing decision is implemented at the MAC layer in the DU.

C. Resource allocation and modulation optimization

We now define an optimization problem whose objective is to dynamically determine the number of resources ($N_{n,k,t}^{\text{rb}}$) and the modulation order ($M_{n,k,t}$) assigned to each active UE of every cell ($k = 1, \dots, K_{n,t}^a$, $n = 1, \dots, N_t^a$), on a given slot t , so that the air interface performance is optimized and the capacity of the shared FH is not exceeded (i.e., $S_t^{\text{FH}} \leq C$). Our objective is to serve all the UEs that have data on each slot and, therefore, a suitable optimization criterion is the *maximization of the minimum per-user throughput*, by considering all the active UEs in the optimization. As the optimization will be performed on a slot-basis, slot index t will be omitted.

If we focus on maximizing the minimum of the per-user throughput ($S_{n,k}^{\text{AI}}$) among all active UEs across the cells, subject to a shared FH capacity constraint, the optimization problem for dynamic modulation compression and resource allocation can be set as follows:

$$\begin{aligned} & \underset{\{M_{n,k}, N_{n,k}^{\text{rb}}\}}{\text{maximize}} && \min_{n,k} \left(\frac{1}{T} N^{\text{sc}} N^{\text{os}} N_{n,k}^{\text{rb}} M_{n,k} \right) \\ & \text{subject to} && \sum_{n=1}^{N^a} \sum_{k=1}^{K_n^a} \frac{1}{T} (N^{\text{sc}} N^{\text{os}} N_{n,k}^{\text{rb}} M_{n,k} + \mathcal{O}) \leq C \end{aligned} \quad (3)$$

where $\mathcal{O} = \mathcal{O}_{\text{mac}} + \mathcal{O}_{\text{mc}}$ is the total overhead for implementing 7.2x intra-PHY split and dynamic modulation compression.

Note that the objective of this work is to analyze solutions for configurations in which the shared FH interface is the one that may saturate the system and consequently constraints the system performance. We are not interested in situations in which the system’s saturation is given by the air interface, because MAC schedulers for this purpose have been largely investigated in literature. Instead, we want to assess the impact that the FH interface has over the air interface performance. Because of such an assumption and focus, we are omitting in (3) the constraint related to the number of RBs that each cell has available for transmission.

For mathematical tractability purposes, we assume that the product of the modulation order and the number of assigned RBs (i.e., $x_{n,k} = N_{n,k}^{\text{rb}} M_{n,k}$) is a continuous optimization variable. Hence, the optimization problem in (3) is a convex optimization problem on $\{x_{n,k}\}$, since the objective function is convex and the constraint is a linear function on $\{x_{n,k}\}$ [12]. As a result, it can be solved using standard convex optimization tools. However, the optimal solution can be found in closed form by reformulating the problem, as detailed next.

The optimization problem in (3) can be reformulated equivalently by including a new optimization variable (γ) that encapsulates the minimum of the per-user throughputs:

$$\begin{aligned} & \underset{\{x_{n,k}\}}{\text{maximize}} && \gamma \\ & \text{subject to} && \sum_{n=1}^{N^a} \sum_{k=1}^{K_n^a} \frac{1}{T} (N^{\text{sc}} N^{\text{os}} x_{n,k} + \mathcal{O}) \leq C \\ & && \frac{1}{T} N^{\text{sc}} N^{\text{os}} x_{n,k} \geq \gamma \quad \forall n, \forall k \end{aligned} \quad (4)$$

The optimization problem in (4) is a linear programming problem with respect to $\{x_{n,k}\}$ and γ , since the objective function is a linear function and the constraints are both linear functions on $\{x_{n,k}\}$ and γ [12]. Interestingly, the optimal solution to (4) is such that the resulting per-user throughput of all users is the same and the shared FH capacity is fully distributed among the users so that both constraints in (4) are met with equality. Said equivalently, the optimal solution lies in the boundaries of the constraint set, as we elaborate next.

In (4), γ is maximized when all the terms $\frac{1}{T} N^{\text{sc}} N^{\text{os}} x_{n,k}$ are equal, i.e., $\frac{1}{T} N^{\text{sc}} N^{\text{os}} x_{1,1} = \frac{1}{T} N^{\text{sc}} N^{\text{os}} x_{n,k}, \forall n, \forall k$. Note that, if such a condition is not met, then we could always tune the solution to increase γ and so maximize the objective function. Therefore, the optimal solution to the optimization problem in (4) satisfies the second constraint with equality and has the following structure:

$$x_{n,k} = \frac{T\gamma}{N^{\text{sc}} N^{\text{os}}}, \forall n, \forall k. \quad (5)$$

In addition, the optimal solution to (4) satisfies the FH capacity constraint (first constraint) with equality. Note that in the case that we have a solution for which the FH capacity constraint is not met with equality, then we could always increase all $\{x_{n,k}\}$, increasing the objective function γ , and so the original solution was not optimal. Therefore, by including the structure in (5) into the FH constraint in (4) with equality, and elaborating with the expression, we can isolate γ as:

$$\gamma = \frac{C - \sum_{n=1}^{N^a} K_n^a \frac{\mathcal{O}}{T}}{\sum_{n=1}^{N^a} K_n^a}. \quad (6)$$

Finally, by combining both expressions in (5) and (6), the optimal solution to the optimization problem in (4) results:

$$x_{n,k}^* = \frac{\left(TC - \sum_{n=1}^{N^a} K_n^a \mathcal{O}\right)}{N^{\text{sc}} N^{\text{os}} \sum_{n=1}^{N^a} K_n^a}, \forall n, \forall k. \quad (7)$$

As the optimization problem in (4) is equivalent to the optimization problem in (3), the solution in (7) is the optimal solution to the optimization problem in (3) as well. However, in the original optimization problem in (3), we had two sets of optimization variables (i.e., the modulation order and the RB assignment per UE). Therefore, multiple solutions for the modulation order and RB assignment result to be optimal. Among the multiple solutions that can be found, in this paper we focus on two practical variants:

- fix the modulation order and limit the RB assignment per UE based on (7), or
- fix the RB assignment and limit the modulation order per UE based on (7).

In both cases, as shown in (7), the actual number of users that are active (with data in their RLC buffers) ($\sum_{n=1}^{N^a} K_n^a$) are required, as well as the available FH capacity (C) and the overheads needed to implement the intra-PHY split and the dynamic modulation compression ($\mathcal{O} = \mathcal{O}_{\text{mac}} + \mathcal{O}_{\text{mc}}$).

In case of limiting the RB assignment per UE, for each UE we can select the modulation order ($M_{n,k}$) based on the

reported CQI from the k -th UE in the n -th cell, and then select the number of RBs for that UE in such a way that (7) is met. That is,

$$N_{n,k}^{\text{rb}}{}^* = \frac{\left(TC - \sum_{n=1}^{N^a} K_n^a \mathcal{O}\right)}{N^{\text{sc}} N^{\text{os}} M_{n,k} \sum_{n=1}^{N^a} K_n^a}, \forall n, \forall k. \quad (8)$$

In case of limiting the modulation order per UE, we can do the resource selection as usual, allocating RBs across the different UEs in a cell, and then constrain the modulation order per UE in such a way that (7) is met. That is,

$$M_{n,k}^* = \frac{\left(TC - \sum_{n=1}^{N^a} K_n^a \mathcal{O}\right)}{N^{\text{sc}} N^{\text{os}} N_{n,k}^{\text{rb}} \sum_{n=1}^{N^a} K_n^a}, \forall n, \forall k. \quad (9)$$

Note that for the sake of numerical tractability, the optimization problem in (3) has been solved by assuming that the optimization variables are continuous. Thus, there is a need to quantize the derived solutions into practical values. A simple quantization process is to take the floor over the possible RB numbers and modulation orders.

D. Modulation order limits

As analyzed in [6], [8], an efficient way to address the air-FH trade-off is by limiting the maximum modulation order that is permitted per cell. This per-cell modulation order limit can still be used even if we adopt dynamic methods, as the ones proposed in the previous sections (dropping, postponing, RB optimization, MCS optimization). In fact, it can be of interest in dropping and postponing strategies, as it may produce less dropping of packets and postponing of scheduling decisions due to the inherent limitation of the required FH throughput with lower modulation orders.

IV. SIMULATION RESULTS

For the performance assessment, we use the 5G-LENA module of ns-3 [9] with the 3GPP spatial channel model developed in [13], compliant with TR 38.901, and the NR-based PHY abstraction model proposed in [14].

A. Deployment scenario

We use a typical hexagonal site deployment. Each site has 3 cells (3 RUs), with 3 uniform planar antenna arrays, each covering 120° in azimuth. We deploy three sites, leading to a total of 9 cells ($N = 9$ RUs). We use an Urban Micro scenario, characterized by an inter-site distance of 200 m, RU antenna height of 10 m, and 30 dBm transmit power. $K_n = 10$ UEs are randomly deployed per cell ($\forall n$). The deployment assumes a frequency reuse of 1; so that all cells transmit in the same frequency band. The operational band is in the 2 GHz region, with 100 MHz channel bandwidth per cell and a PRB overhead of 0.04 (typical of NR). Numerology 1 is used, i.e., 30 KHz subcarrier spacing. The duplexing mode is TDD, with dynamic downlink-uplink slot structure. Therefore, the slot length is $T = 0.5$ ms, and the number of RBs and symbols available for downlink data within it are: 266 RBs and $N^{\text{sybm}} = 12$ symbols (2 OFDM symbols are left for PDCCH and PUCCH

control channels), $\forall n, \forall k$. The antenna array configuration consists of 5×2 directional antenna elements at RUs and 1 isotropic antenna element at UEs. The base MAC scheduler uses Round Robin scheduling metric to initially sort the UEs and we configure 1 RB Group = 4 RBs.

Regarding the FH, we assume the scenario shown in Fig. 1. In this way, CUs/DUs are centralized, giving service to all the cells (RUs) of the deployment scenario under consideration, following functional split Option 7-2x (DU-RU). Due to the shared FH topology, the FH capacity is shared among the 9 RUs. We assume an available FH capacity of C Gbps, the value of which will be varied through the simulations, with $C = 1$ Gbps and $C = 0.5$ Gbps. For the required FH throughput, we use the expression in (1) with $\mathcal{O}_{\text{mac}} = 100$ Kbits and $\mathcal{O}_{\text{mc}} = 100$ bits.

B. Traffic model and KPIs

We use File Transport Protocol (FTP) Model 1, as recommended by 3GPP and described in [15, Sec. A.2.1.3.1]. This traffic model can be tuned to capture a range of realistic traffic types [16]. Among these, we have selected a widely used and predominant traffic in the Internet, i.e., video traffic. YouTube video traffic is characterized by a fixed file size of 50 KBytes and $\lambda = 50$ files/second are generated according to a Poisson process, so that in average one file is generated every 20 ms [16]. The simulation duration is 10 seconds. We evaluate the performance of the FTP traffic model over UDP transport protocol. As for the link layer, we consider UDP over RLC-UM.

The key considered end-to-end performance metrics in this study are “user-perceived throughput” (UPT) and “latency”. Both are measured at the IP layer. The user-perceived throughput is computed as the fraction of received bytes per file over the receive time duration to complete the file transfer. The latency is computed as the sum of the delay to receive all packets that compose the file.

C. End-to-end results

In the end-to-end evaluation, for each available FH capacity (C), we assess the impact of using different FH-aware scheduling strategies for dynamic modulation compression. We compare the following methods:

- Drop: Dropping of MAC PDUs at the high-PHY layer, described in Sec. III-A.
- Postpone: Discarding MAC scheduling decisions in the MAC scheduler, detailed in Sec. III-B.
- Rb: Optimization of the RB assignment per active UE, as per (8).
- Mcs: Optimization of the modulation order (and MCS) assignment per active UE, as per (9).

In all cases, we consider two subvariants of the methods: 1) no modulation order per cell limitation (i.e., each cell can use up to 256QAM) and 2) limiting the modulation order per cell to 16QAM, which are depicted in the legends of the figures as +256QAM and +16QAM, respectively.

Fig. 2. End-to-end performance metrics, for different FH-aware scheduling strategies. (a) UPT CDF (Mbps), (b) latency CDF (ms). $C = 1$ Gbps.

Fig. 2 and Fig. 3 show the Cumulative Density Function (CDF) of the obtained end-to-end performance metrics ((a) UPT (in Mbps) and (b) latency (in ms)) for the cases of $C = 1$ Gbps and $C = 0.5$ Gbps, respectively. Results are displayed for the different scheduling strategies separately (Drop, Postpone, Rb, and Mcs), combined with MCS limits (256QAM and 16QAM).

In the case of $C = 1$ Gbps, we can observe that Postpone, Mcs, and Rb strategies perform pretty similar in terms of UPT and latency (see Fig. 2). Interestingly, they outperform the Drop strategy when no modulation order limit is imposed (i.e., +256QAM in figures), which is the simplest implementation strategy but has the drawback that it drops and loses data packets, as shown by the tail of the UPT and latency CDFs. In case the 16QAM modulation order limit is imposed, all the strategies perform similarly, because, for $C = 1$ Gbps, the end-to-end performance is limited by the air interface (available RBs) and not the FH.

These conclusions are not valid though, if we have a FH of lower available capacity, as with $C = 0.5$ Gbps, shown in Fig. 3. In this case, we can see a clear advantage of the FH-aware scheduling strategies that optimize the RBs and MCS

Fig. 3. End-to-end performance metrics, for different FH-aware scheduling strategies. (a) UPT CDF (Mbps), (b) latency CDF (ms). $C = 0.5$ Gbps.

allocation (Rb and Mcs) over the other strategies (Drop and Postpone) when no modulation order limit is imposed (i.e., +256QAM in figures). In particular, Rb and Mcs optimization strategies provide a significant improvement gain over the Postpone strategy (see Fig. 3(a)) in all percentiles of the UPT and latency. Among Rb and Mcs strategies, we can observe that the Mcs strategy slightly outperforms the Rb strategy in the 5%tile of the UPT, because when we reduce the modulation order, we are making the transmission more robust towards channel impairments and interference conditions (the probability of error is reduced). On the other hand, the Rb strategy outperforms the Mcs strategy in the 50%tile UPT and 95%tile UPT (see Fig. 3(a)), because it allows to distribute all the available air interface RBs more effectively, enabling to serve a larger quantity of data. Interestingly, in case a modulation order limit is imposed (+16QAM in figures), the UPT values are upper bounded but the 5%tile and 50%tile of the UPT can be improved, as compared to its counterparts with no MCS limit, because the reduction of the modulation order improves the transmission reliability.

V. CONCLUSIONS

This paper has proposed dynamic holistic and optimized FH-aware scheduling strategies for hybrid C-RAN architectures with multiple cells sharing a single FH link, assuming that modulation compression is used to reduce the amount of data that is sent over the FH interface. The proposed optimized methods allow controlling the FH resources' utilization without exceeding the available FH capacity while optimizing the per-user experienced throughput of all the users in the system. We have then evaluated the proposed FH-aware scheduling methods in an end-to-end, dynamic, multi-cell simulation built on the ns-3 5G NR module. Simulation results in a multi-cell and multi-user scenario have shown that the proposed optimized solutions outperform the baseline strategies, providing user-perceived throughput gains in conditions of a tight FH capacity. In situations with a more relaxed FH capacity, simulations show that the baseline holistic method that follows a postponing approach is already sufficient from a system-level performance perspective.

REFERENCES

- [1] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent advances in cloud radio access networks: System architectures, key techniques, and open issues," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 2282–2308, 2016.
- [2] 3GPP TS 38.401, *TSG RAN; NG-RAN; Architecture description*, Release 16, v16.1.0, Mar. 2020.
- [3] O-RAN Fronthaul Working Group, *Technical Specification; Control, User and Synchronization Plane Specification*, v03.00, Apr. 2020.
- [4] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A survey of the functional splits proposed for 5G mobile crosshaul networks," *IEEE Commun. Surveys Tutorials*, vol. 21, no. 1, pp. 146–172, 2019.
- [5] S. Parkvall *et al.*, "NR: the new 5G radio access technology," *IEEE Commun. Standards Mag.*, vol. 1, pp. 24–30, Dec. 2017.
- [6] S. Lagen, L. Giupponi, A. Handsson, and X. Gelabert, "Modulation compression in next generation RAN: air interface and fronthaul trade-offs," *IEEE Commun. Magazine*, Jan. 2021.
- [7] A. Umesh, T. Yahima, T. Uchino, and S. Okuyama, "Overview of O-RAN Fronthaul Specifications," *Special Articles on Standardization Trends towards Open and Intelligent Radio Access Networks*, NTT Docomo technical journal, vol. 21, July 2019.
- [8] S. Lagen, X. Gelabert, L. Giupponi, and A. Handsson, "Semi-static modulation compression optimization for next generation RANs," *2021 IEEE Int. Conf. on Commun.*, June 2021.
- [9] N. Patriciello, S. Lagen, B. Bojovic, and L. Giupponi, "An E2E simulator for 5G NR networks," *Elsevier Simulation Modelling Practice and Theory*, vol. 96, p. 101933, 2019.
- [10] O-RAN Alliance White Paper, *O-RAN use cases and deployment scenarios*, Feb. 2020.
- [11] 3GPP TR 38.801, *TSG RAN; Study on new radio access technology: Radio access architecture and interfaces*, Release 14, v14.0.0, Mar. 2017.
- [12] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, 2004.
- [13] T. Zugno, M. Polese, N. Patriciello, B. Bojovic, S. Lagen, and M. Zorzi, "Implementation of a spatial channel model for ns-3," *ns-3 Workshop 2020*, June 2020.
- [14] S. Lagen, K. Wanuga, H. Elkotby, S. Goyal, N. Patriciello, and L. Giupponi, "New radio physical layer abstraction for system-level simulations of 5G networks," in *IEEE Int. Conf. on Commun.*, June 2020.
- [15] 3GPP TR 36.814, *Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer aspects*, Release 9, v9.2.0, Feb. 2017.
- [16] "Application of stochastic geometry in modeling future LTE-A and 5G wireless networks," https://web.ma.utexas.edu/conferences/simons2015/Modeling3GPP_Samsung_Final.pdf. Samsung.